图书情报工作
第63卷 第19期 2019年10月

新一代机构知识库的关键技术和发展趋势研究*

■ 崔海媛 孙超 罗鹏程
北京大学图书馆 北京 100871

摘要：［目的/意义］通过调研国内外新一代机构知识库的研究现状和服务需求，分析关键技术和功能特点，提出发展趋势，为建设新一代机构知识库发展提供建议。［方法/过程］通过文献研究对机构知识库技术和功能发展趋势进行总结，并根据相关研究成果，分析新一代机构知识库的11个关键技术、标准和协议。最后通过研究和实践经验提出新一代机构知识库在框架、功能、服务目标等方面的发展趋势。［结果/结论］新一代机构知识库发展趋势包括：从机构学术仓储到机构信息基础设施；从自存档到自动提交；从独立平台到与科研管理系统融合与发展；从学术成果管理平台到学术资源服务中心；从学术成果数据检索到大数据语义研究支持；从成果存档到新型学术交流社区；从应用计量指标到建立全新学术评价体系。

关键词：机构知识库 新一代机构知识库 下一代机构知识库 关键技术 发展趋势
分类号：G251.7
DOI：10.13266/j.issn.0252-3116.2019.19.009

1 引言

开放获取运动推动下，全球机构知识库（Institutional Repository，IR）发展迅速，至2019年3月，在机构知识库注册网站（The Directory of Open Access Repositories，OpenDOAR）注册的机构知识库数量从2005年12月的88个增加到3 996个[1]。然而与开放获取理念的广泛接受、机构知识库数量快速增长和全球开放获取运动推动者的不懈努力相比，机构知识库在学术交流系统中仍未发挥出期望的价值，商业数据库和商业出版依然是学术交流的主要渠道，开放获取打破学术交流商业垄断的目标仍未实现，全球IR的质量和影响力仍存在巨大差距。2016年，IR最早建设者与引领者麻省理工学院，庆祝自该校OA政策发布以来，IR教师发表论文存储率达到44%。同年，俄勒冈州立大学（Oregon State University）和内布拉斯加州立大学（Nebraska State University）的IR存储率超过40%。2017年3月，美国大学与研究型图书馆协会（Association of College and Research Libraries，ACRL）发布《2017环境扫描》报告[2]，其中关于机构知识库的发展，报告特别提出，美国大学机构知识库一直存在较低存储率的情况。美国大学IR长期以来，存储率一直低于50%。加州大学（The University of California）IR储存率2016年仅为25%，没有存缴政策支持的其他美国大学IR存储率则更低。而学术交流环境，却在大数据、云计算、泛在网络、虚拟现实、人工智能等新技术广泛应用影响下，已经发生改变。新学术交流生态环境下，IR如何发挥作用，全球IR是否能通过新技术形成合力，IR如何迎接新挑战和新机遇，新一代机构知识库的研究和构建成为必然选择。

本文将调研国内外新一代机构知识库的研究现状和服务需求，分析新一代机构知识库关键技术，提出新一代机构知识库功能特点，为建设新一代机构知识库发展提供建议。

2 研究综述

2.1 机构知识库功能与服务增强研究与实践

对IR功能扩展和提供增值服务的研究和实践，国内外一直都有积极的探索。在中国，马建霞[3]提出了机构知识库在内容建设与服务设计方面的趋势，如制定强制存缴政策、采取灵活的访问策略、简化存缴步骤、集成到用户信息环境、以机构知识库联盟获得规模

* 本文系北京大学桐山教育基金研究资助项目"中日韩研究论文开放获取与学术影响力数据分析与研究"研究成果之一。
作者简介：崔海媛（ORCID：0000-0001-5541-7100），副研究馆员，硕士，E-mail：cuihy@pku.edu.cn；孙超（ORCID：0000-0002-3180-2669），馆员，硕士；罗鹏程（ORCID：0000-0001-9598-0715），馆员，硕士。
收稿日期：2019-01-10 修回日期：2019-03-13 本文起止页码：96-104 本文责任编辑：杜杏叶

崔海媛,孙超,罗鹏程.新一代机构知识库的关键技术和发展趋势研究[J].图书情报工作,2019,63(19):96-104.

优势、提供知识审计与能力分析功能、长期保存服务、技术和服务团队的持久保障等。张晓林[4]针对机构知识库的发展,提出机构知识库支持非文本信息存储利用、支持教育科研活动、支持机构战略性知识管理三个未来发展趋势以及一系列可能的服务功能。刘巍、祝忠明、吴志强等[5-8]研究内容可视化知识图谱,图像检索、影音资源支持、检索三维模型等新技术在机构知识库中的应用,提供增值服务。张旺强等[9]通过互操作协议,实现简化用户提交过程,自动存档。崔海媛等[10]通过提供增强内容和数据统计,设计更符合资助机构应用的机构知识库。香港科技大学机构知识库以学者为中心,展示学者成果,并利用可视化技术,构建学者的合作者网络,将学者的 Scopus ID、Researcher ID 和 ORCID 三者进行关联,全方位展示学者的学术轨迹[11]。香港大学机构知识库将论文与学者、基金项目与学者进行了关联[12]。香港理工大学机构知识库在论文的页面详细展示了论文的计量指标——Scopus 的被引频次、WoS 的被引频次、访问次数、下载次数、Altmetric 等信息[13]。

国际上,对 IR 的功能拓展和服务增强研究更为广泛、深入。L. Sterman 等[14]用丰富的可视化工具,提供增强的统计和计量数据服务,并基于访问情况,主动向作者推送信息服务。A. Cocciolo[15]研究比较 Web2.0 应用对用户参与 IR 的不同,发现应用 Web2.0 技术有助于增加用户对 IR 的兴趣与参与。J. Richard[16]提出互操作如何促成机构知识库的发展、可能被使用的方式以及实现的方式。机构知识库与科研管理系统融合研究与实现成为趋势,香港大学的科研管理系统是在 IR 基础上扩展实现,通过在 DSpace 的基础上增加 CRIS 模块,实现了 CERIF 兼容的 DSpace-CRIS 系统[17]。伦敦国王学院(King's College London)[18]、加拿大皇后大学(Queen's University)[19]在科研管理系统的基础上,增加具有开放获取功能的机构知识库模块。圣安德鲁斯大学(Universtiy of St Andrews)的机构知识库与科研管理系统通过 API 的方式实现系统间互操作[20]。

2.2 新一代机构知识库相关研究

2016 年 4 月,开放获取知识库联盟(The Confederation of Open Access Repositories, COAR)成立下一代知识库工作组,调研用户新需求,提出新功能和技术方案。COAR 认为,利用分布在全球的 3 000 多个知识库网络,创建更加可持续创新的系统,共享研究成果,可以提供全球研究的全面视野,同时也使得每个学者和机构都能够参与全球的科学和学术研究网络。在知识库网络之上建立标准化使用统计指标、同行评议和社交网络等附加服务,将有利于机构知识库发展,改变商业出版商主导学术传播体系的现状[21]。

2017 年 11 月,COAR 发布工作组研究成果报告“下一代系统-行动和技术建议”,介绍工作组研究成果,提出对下一代机构知识库应用新技术、标准、协议的建议,以帮助知识库融入网络环境,使它们在学术交流生态系统中发挥更大作用。下一代 IR 的建设目标是:使知识库成为分布式的全球网络化学术交流体系的基础,在此之上提供各种层级的增值服务,从而变革出版系统,使其更加以研究为中心、更加开放并支持创新,同时由学术界共同管理。这一愿景的一个重要组成部分,是知识库将提供多种研究成果的开放获取,支持学术成果广泛传播,并使其在研究评估过程中得到正式承认。报告描述了 11 项新功能,以及在知识库现有功能基础上开发包括社交网络、同行评议、通知和使用统计等新服务相关的技术、标准和协议,包括:①公开标识符;②在资源层声明许可协议;③通过导航发现;④与资源互动(注释、评论和评述);⑤资源转移;⑥批量发现;⑦收集和导出活动元数据;⑧用户识别;⑨用户认证;⑩公开标准化使用计量指标;⑪长期保存[22]。

在 COAR 引领下,对新一代机构知识库的功能、技术和发展的研究和应用,成为 IR 领域的热点。

2018 年 9 月 4 日,来自法国、英国、荷兰、意大利等 11 个欧洲国家的主要科研经费资助机构,在欧洲研究委员会(European Research Council, ERC)的支持下,联合签署了新的开放获取计划——cOAlition S 计划(以下简称“S 计划”)。S 计划的核心原则是:“从 2020 年 1 月 1 日起,所有由上述 11 国以及欧洲研究委员会拨款支持的科研项目,都必须将研究成果发表在完全开放获取期刊或出版平台上。” S 计划作为 OA2020 的行动计划,带动全球加快开放出版步伐,改变传统学术出版格局[23]。2018 年 11 月,惠康基金会和比尔及梅琳达·盖茨基金会加入 S 计划。惠康基金会和比尔及梅琳达·盖茨基金会更新了开放获取政策:2020 年 1 月起,资助项目成果论文全部需要开放获取,可以从 PMC 和 EuropePMC 检索,并将不再承担受资助者在混合开放获取(同时拥有订阅和免费内容)期刊上发表文章的费用[24]。2018 年 12 月 2-4 日,在德国召开了第 14 届开放获取柏林会议,37 个国家的资助机构、科研与教育机构和图书馆参加,进一步协调推进立即全面开

放获取的政策。与会代表一致同意:努力保证作者保留论文著作权,努力实现论文的全面立即开放获取,努力建立短期的过渡性的转型协议将订购期刊转换为开放出版,而且这些协议至少应不增加成本并在以后随着市场转换而调整,会议代表期待出版社与国际科研界一起共同努力实现论文的全面立即开放获取。中国国家自然科学基金委、国家科技图书文献中心、中科院文献情报中心代表在会议上发布立场声明,明确表示中国支持 OA2020 和开放获取 S 计划,支持公共资助项目研究论文立即开放获取[25]。

S 计划在实施建议中,要求 IR 必须在 OpenDOAR 中注册或正申请注册。此外,还需要遵循以下标准:①提供自动存缴功能;②按照 JATS 等标准以 XML 格式保存全文;③以标准互操作格式提供高质量的元数据,包括关于出版物 DOI、缴存版本(AAM 或 COR)、开放获取状态、缴存版本许可等内容;④符合 cOAlition S 元数据标准规范;⑤提供开放 API,允许其他人(包括搜索引擎等机器)对内容进行访问;⑥提供质量保障机制来将内容全文与核心的文摘索引服务系统关联起来(例如 PubMed);⑦ 保证长期可靠运行;⑨提供帮助咨询服务[26]。

S 计划已经影响机构知识库的发展方向。COAR 响应 S 计划,将会在下一代机构知识库技术规范中,支持 S 计划。但同时对 S 计划部分技术标准提出意见和修改建议,如:①自动存缴解决方案尚不成熟,不应强制要求;②XML 格式过于耗费资源,应要求符合获取全文的标准规范,如符合 Signposting 协议,而非强制要求存储为 XML 格式。③开放 API 要求过于模糊,不便于操作,应提供部分推荐 API 建议。④应增加允许 OpenAIRE 收割数据要求。(欧洲开放获取基础设施研究项目,Open Access Infrastructure Research for Europe,简称 OpenAIRE)。⑤帮助咨询服务是大多数机构网站提供的服务,不必做为 IR 的强制要求[27]。正在开发并计划在 2019 年发布的 DSPACE7.0 版本,已经在根据 S 计划,增强自动存缴、支持 XML-JATS、更多样的 API 服务等功能[28]。

3　新一代机构知识库技术和功能特点

新一代机构知识库突破"数字化采集和完整保存单一或多个大学的知识成果产出的交流平台"[29]的机构知识库定义,将机构知识库定义提升为:分布式全球网络化新型学术交流体系的基础,以研究为中心、提供面向研究所需的开放增值服务,提供多种研究成果的开放获取,支持学术成果广泛传播,推动建立新型学术出版和研究评估的平台。

重新思考开放获取,突破现有机构知识库存档、管理、发布、检索和开放共享的目标定位,打破单一机构限制和学者服务范围,从全球学术交流和创新研究视野出发,面向未来,新一代机构知识库的建设目标是成为全球新型学术交流生态系统的重要基础设施,能够管理多种类型的学术资源,包括:论文、图书、报告、数据、软件、工具等,成为全世界学者学术交流提供服务的学术资源中心。新一代机构知识库是能够实现数据互操作的网络知识库,是用户友好、机器友好的知识库,是为每个学者和机构提供设施、数据和服务的全球学术交流基础。在新一代机构知识库上,能够实现多种增值服务,包括学术评价、同行评审和学术社交等,以学术交流为中心,全面支持学术开放和研究创新。

3.1　新一代机构知识库的功能、技术、标准和协议

COAR 提出新一代机构知识库应具有的 11 种技术、标准和协议。图 1 是新一代 IR 与现有 IR 的架构对比图[22],在新一代机构知识库中,通过集成云计算、搜索和内容管理等全新技术,设计全新 IR 云基础设施,提供更多服务协议、标准和服务,为 IR 开发更多增值服务提供支持。

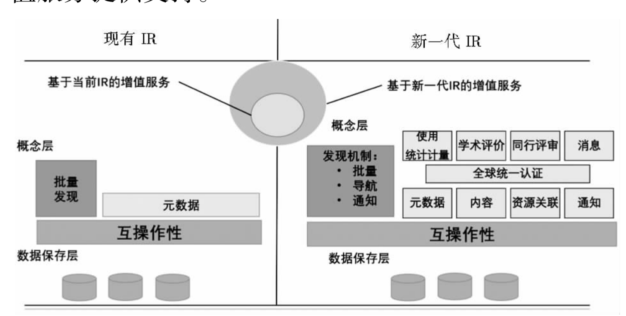


图 1　新一代 IR 与现有 IR 的架构对比

(1)公开标识符(Exposing Identifiers)。访问 IR 等学术门户网站时,用户可以轻松找出目标网页、书目记录链接、作者身份等。但是,由于门户网站使用不同的方法展示内容,对搜索引擎等收割数据服务却很难解决。如何能够方便用户定位与引用 IR 资源,如何能够让数据交互更为顺畅,让搜索引擎能够准确获取数据,IR 中的数据(元数据和成果)具有唯一标识是可行的解决方案。

Signposting 是一种使学术网络对机器更友好的方法。它使用 Typed Links 方法来区分学术门户中重复出现模式。对于任何媒体类型的资源,类型链接都会在 HTTP 链接头中提供。对于 HTML 资源,它们还在

HTML 链接元素中提供。Signposting 使用类型链接(在 HTTP 链接头,HTML <link> 元素或 <rs:ln> ResourceSync 元素中)来判断学术门户重复内容出现模式。Signposting 可用于支持自动发现与学术对象有关的各种资源,包括书目描述、持久标识符、许可信息、作者或作为对象一部分的各种资源。在网站上采用 Signposting 方法,能够允许机器以统一的方式定位学术门户内容,有助于数据互操作。HTTP 链接头方式有很多好处,标头方法可用于任何媒体类型的资源,而不仅仅是 HTML。因此,图像、数据集、PDF 等都可以统一使用相同的方法来阐明模式。可以使用仅返回事务元数据而非内容的 HTTP HEAD 请求来访问标头,可以获得大量资源的头部,例如大数据集或高分辨率图像,而无需实际下载这些资源。以类似的方式,HTTP HEAD 请求可用于获取受限内容的标头,包括付费应用文章[30]。

(2)在资源级别声明许可协议(Declaring Licenses at a Resource Level)。如何让用户与机器都能清楚 IR 资源的知识产权许可情况?在内容组织中,增加明确的许可标识,在 HTTP 链接增加许可信息,是有效的解决方法。通过 Signposting 方法,在 HTTP 链接增加知识产权协议,如 Creative Commons Copyright Licenses 相关内容,是解决方案之一。

(3)通过导航发现(Discovery through Navigation)。IR 中的成果数据类型丰富,一条元数据可能会对应 PDF 和/或 HTML 版本的论文、一个或多个支持数据集、图书或表格附件等。为了帮助机器准确识别数据对象,实现准确的搜索和导航,在 HTTP 链接中提供数据链接关系、数据类型等是有效的解决方法。通过 Signposting 方法,在 HTTP 链接中增加一组 Web 资源信息是一种解决方案。

(4)与资源交互(注释、评论和评述)(Interacting with Resources (Annotation, Commentary and Review))大量研究已经证明,提供用户交互功能,能够增加用户参与度。通过接入第三方社交媒体服务,允许用户注释、评论和评述,IR 可以发挥学术交流中心作用,促进研究者讨论和协作工作。

Activity Streams 2.0 是一种描述与资源交互的方法,包括评论、点赞、共享等。交互表示为 JSON-LD 并使用 Activity Streams 2.0 词汇表。虽然该词汇表针对的是一般社交网络活动,但仍可以扩展学术词汇表[31]。Web Annotation Model and Web Annotation Protocol 是专门表达注释(包括评论、评述等)的方法以及用于创建和管理它们的相关协议。注释使用基于 RDF 的词汇表来表达,并且可以呈现为 JSON-LD。该协议基于 HTTP 并遵循 REST 设计原则[32]。国际图像互操作性框架(The International Image Interoperability Framework,IIIF)是一个支持图像互操作 API 的协议,用于图像复用、共享和与交互。应用 IIIF 协议,可以对图像实现操作、评论、引用、分享和认证访问等功能[33]。

(5)资源转移(Resource Transfer)。分布式、网络化、云存储模式,是新一代机构知识库的核心架构,需要实现资源内容的分布式部署。云存储和云计算技术已经成熟,能够支持所需应用实现。

IPFS 是点对点超媒体协议,旨在使网络更快、更安全、更开放。应用 IPFS 协议,可以实现多方之间共享大数据集合的需求[34]。ResourceSync 是一种基于站点地图的规范,存储库管理器可以使用该规范提供信息,允许第三方系统持续与其存储库中的资源保持同步,即创建、更新和删除。站点地图允许公开知识库内容和搜索引擎所需的元数据。ResourceSync 可用使用 Sitemaps XML 格式实现内容和元数据的发现和同步[35]。SWORD(简单知识库内容 Web 服务提供)是一种轻量级协议,用于将内容从一个位置存储到另一个位置[36]。

(6)批量发现(Batch Discovery)。随着 IR 的发展,用户需要统一、跨平台的知识库资源发现服务,需要资源文本内容的搜索。实现全球知识库学术搜索功能,是新一代机构知识库的重要目标。使用 ResourceSync, Signposting, Sitemaps 等协议应用,实现批量搜索,能够提升知识库资源价值。Sitemaps 提供了易于搜索引擎抓取网站内容的方法。在最简单的形式中,Sitemap 是一个 XML 文件,其中列出了每个可用资源的 URL 以及有关该资源的可选附加元数据(例如修改日期、更改频率等),有助于爬虫及时、准确获取数据[37]。

(7)收集和公开活动(Collecting and Exposing Activities)。机构知识库需要主动并实时收集和公开活动(包括任何修改、增加、评论、注释、同行评议、访问、下载等),并实时发送通知给相关用户,提供用户所需的多种增值服务,使 IR 成为学术交流社区。实现通知机制,除了需要资源对象具有唯一标识符和用户需要身份认证外,还需要应用多种标准协议和技术。

Activity Streams 2.0 为资源活动信息提供语义定义规范,通过 JSON 格式和词表规范,提供活动结构化描述方法。关联数据通知(Linked Data Notifications)是

一种通用通知协议,描述服务器(接收方)如何将应用程序(发件人)推送给它们的消息,以及其他应用程序(消费者)如何检索这些消息。任何资源都可以通知消息的接收端点(收件箱)。消息以 RDF 格式定义,可以包含任何数据。其中任何资源都可以通知收件箱,该收件箱可以发布与该资源相关的通知。例如,注释、评论或审阅信息,通知该资源发生的交互、交互内容、交互参与者等。通知表达方式为 JSON-LD 并使用 Activity Streams 2.0 词汇表[38]。ResourceSync Change Notifications 是基于 WebSub 的发布/订阅协议,并向订阅者发送知识库资源相关修改(创建/更新/删除)的通知。ResourceSync 通知可用于内容和元数据的发现和同步,并使用 Sitemaps XML 格式[39]。Webmention 是一种点对点的 trackback/pingback 方法,旨在通知资源链接变化,支持双向链接[40]。WebSub 是一种出版/订阅协议,出版者将资源更新通知发布订阅用户。机构知识库通过 WebSub 与出版社实现资源交互,及时获取论文引用、评论、评述等数据[41]。其他消息传递协议(例如 AMQP、Kafka)为所有 Web 内容发布者和订阅者交互提供了通用通信机制。

(8)用户识别(Identification of Users)。资源交互和活动公开等所有增值功能和服务,都需要用户具有唯一身份标识,需要识别资源与用户的关联关系。身份标识可以使用 ORCID、Social Network Identities、WebID 等。ORCID(Open Researcher and Contributor Identifier,开放研究者与贡献者身份标识符)提供一个永久性的数字标识符给研究者,并通过与主要研究工作流程(例如手稿和出版成果)集成,实现研究者与学术活动的自动链接,识别研究者学术成果[42]。社交媒体身份标识(Social Network Identities)由多个社交网络平台提供。WebID 是一个代理 HTTP(S) URI,通常由代理(个人、组织、设备等)在所属域中创建。WebID 基于 RDF 的配置文件机器可读,通常与 WebID/TLS 身份认证和 Web 访问控制认证方法结合使用[43]。

(9)用户认证(Authentication of Users)。提供用户交互与个性化增值服务,需要用户身份识别和认证功能,包括学术身份(如 ORCID)和社交网络(如 Twitter、Google、Facebook、微博、Mastadon)身份。

HTTP 签名提供了类似于 WebID/TLS 的身份验证方法。Sign HTTP messages 除了身份验证之外,它还允许验证客户端和服务器之间的通信未被篡改。该方法目前正在 IETF 申请标准,值得进一步关注[44]。OpenID Connect 1.0 是在 OAuth 2.0 协议之上的简单身份层,用于分布式身份验证。OpenID Connect 允许客户端应用程序(例如机构知识库和浏览器)通过用户身份提供者进行身份验证。认证成功后可以将关于用户的基本信息返回给客户端应用程序。该协议支持可扩展,允许开发应用者使用可选功能,如:身份数据加密、OpenID 服务方信息和 Session 会话管理。主要社交媒体已经支持 OpenID Connect,ORCID 目前在测试阶段[45]。WebID / TLS 是基于传输安全层协议(TSL),X.509 证书和 WebID 等实现安全用户身份验证的协议。它使用户只需从浏览器给出的证书中选择所需证书即可进行身份验证,用于解决服务器获取用户私钥信息和用户 WebID。通过 WebID,获取包含用户私钥的个人信息并进行验证。WebID / TLS 虽然是完全分布式的高效方法,但由于难以生成证书和用户界面,一直没有得到广泛应用[46]。

(10)公开标准化使用计量指标(Exposing Standardized Usage Metrics)。通过共享用户交互数据,机构知识库可以开发和提供更多用户需要的增值服务。收集、管理和提供标准使用计量指标数据,是能够让作者和所有用户了解机构知识库价值所在的重要服务。为保证数据准确、可靠、可信,需要采用通用标准协议、方法和互操作,让用户看到完整的计量数据。如果能够基于数据,建立全球机构知识库标计量指标体系,提供与商业期刊无关的评价系统,意义将更为深远。通过定量数据和用户交互(注释、评论、评述)的定性数据结合应用,机构知识库有可能完成这样的目标。定量数据可以通过两种模式实现:获取模式(如使用 SUSHI)或推送模式(如:google - analytics, IRUS - UK, OpenAIRE 使用的 Piwik,RAMP)。公开使用指标需要解决公开障碍,需要通过通用标准推动,而不仅仅是技术。COUNTER 标准使用户能够获取电子资源的使用统计。该标准被称为“行为准则”,确保供应商和出版商能够为用户提供一致、可靠和可比较的使用数据[47]。SUSHI 是 ANSI / NISO 标准,它定义了用于收割电子资源使用数据的自动请求和响应模型,与 COUNTER 一起使用。ETag 或 entity tag 是 HTTP 的一部分,它是 HTTP 为 Web 缓存验证提供的几种机制之一,它允许客户端进行条件请求。这允许缓存更高效并节省带宽,因为如果内容未更改,则 Web 服务器不需要发送完整响应。ETag 还可用于并发控制,作为一种防止资源同时更新导致互相覆盖的方法,有助于支持系统仅获取有关指标的新数据[48]。

(11)资源长期保存 (Preserving Resources)。开放

获取的意义不仅在于开放访问现在的学术资源,还在于永久访问和长期保存。长期保存不需要每个知识库独立进行,而是应该通过标准、协议和互操作,建立机构和全球学术资源的长期保存网络。保存,需要保持资源(资源、元数据和结构信息) 的复杂互连,还需要通过新技术实现实时获取和保存数据,数据格式应尝试应用可重复使用的格式(如 Latex 和 TEI,而不是 PDF)。长期保存是一项极其复杂的活动,涉及政策、标准、实践和技术等,需要重视、研究和应用。

3.2 新一代机构知识库的发展趋势

3.2.1 从机构学术仓储到机构信息基础设施 新一代机构知识库的建设目标,是成为全球新型学术交流生态系统的重要基础设施,能够管理多种类型的学术资源管理,包括:论文、图书、报告、数据、软件、工具等,成为全世界学者学术交流提供服务的学术资源中心。新一代机构知识库是能够实现数据互操作的网络知识库,是用户友好、机器友好的知识库,是为每个学者和机构提供设施、数据和服务的全球学术交流基础。其设计,以完整学术资源数据汇集为基础,建设学术资源管理与服务平台,实现学术资源管理与服务功能。从而可以实现:改善全球学术交流信息基础设施与环境,促进交叉学科交流与合作;为全世界的研究者提供完整、丰富、多样、创新的学术信息;为研究者提供开放学术信息环境,成为国际一流学术交流的一部分;参与到全球学术交流重构进程中,成为新型学术交流规则的制定者。

新一代机构知识库具有以下功能特点:①基于开放框架设计的云平台架构,提供多种开放服务接口,能够与全球学术交流信息基础设施实现数据互操作和合作服务。②支持多种类型资源的管理和服务,包括学术成果、档案资料、研究数据、软件工具等。③遵循公开标识符标准规范,所有数据都能够基于标识符与全球机构知识库实现互操作。④提供学术资源长期保存、管理与服务,遵循 OAIS 框架,支持公开标识符,提供数据格式进行识别和迁移转换服务。⑤提供统一、完整的学术搜索服务。通过开放服务框架和接口,新一代机构库要么可以提供全新的比现有学术发现更优的发现服务,要么与更优的学术发现合作,为用户提供统一、跨平台的知识库资源发现服务,实现全球知识库学术搜索功能。

以北京大学机构知识库为例,截至 2018 年 12 月,该机构知识库收集了北京大学自 1949 年以来 54 万元数据和 30 万全文数据,逐步建立完整的机构学术论文数据。在数据建设基础上,建设科研管理系统成果子系统,成为机构学术成果仓储。未来将进一步改进和增强功能服务,实现机构学术成果信息基础设施的建设目标。

3.2.2 从自存档到自动提交 新一代机构知识库,通过互操作技术、与出版商合作等方法,从自存档方式转为自动提交工作流程,提供自存档、自动收割数据、代理提交、跨系统合作等多种存缴方式。实现多种提交功能:①通过互操作,与数据库商合作,实现数据自动提交。或与出版商谈判,由出版商直接提供内容。②提供数据工具和接口,嵌入到用户工作流程中,减少自存档操作步骤。如基于数据互操作协 SWORD,实现从 Word 直接提交。数据工具可以自动抽取元数据或者自动生成元数据,改进元数据质量。③全球机构知识库之间数据交互。以美国自然科学基金委员会(National Science Foundation, NSF)的机构知识库(NSF Public Access Repository, NSF-PAR)为例,NSF-PAR 与能源部(the Department of Energy, DOE)和科学和技术信息办公室(Office of Science and Technical Information, OSTI)的机构知识库建立互操作性。2018 年春季开始,两个机构(NSF 和 DOE)资助的出版物的作者可以一次性存放其原稿的最终版本,在 DOE/OSTI 系统中成功存放符合条件的出版物的作者现在可以通过 NSF-PAR 无缝地发布他们的出版物 Pubmed 与 NSF 交换数据[49]。

3.2.3 从独立平台到与科研管理系统融合与发展 实践证明,与科研管理系统融合,机构知识库更具有可持续,能够更好地支持机构学术和科研管理活动,实现新一代机构知识库成为信息基础设施的目标。与科研管理系统融合,机构知识库提供学术资源数据的收集、保存、管理和服务,为研究者提供自动提交服务,为科研管理提供服务。而科研管理流程和研究者确认学术成果数据过程,支持机构知识库完成数据确认流程,得到完整、准确的成果数据,并建立成果和学者的关联关系,建设高质量学术资源仓储,为提供更多增值服务、实现成为学术交流社区和建立全新学术评价体系学术评价奠定了基础。

融合方式可以有多种形式,如:①基于机构知识库的融合。即在机构知识库的基础上进行扩展,增加科研管理的功能。目前,香港大学采用了这种实现方式,通过在 DSpace 的基础上增加 CRIS 模块,实现了 CERIF 兼容的 DSpace-CRIS 系统。②基于科研管理系统的融合。即在科研管理系统的基础上,增加具有开放获取功能的机构知识库模块。目前,伦敦国王学院(King's College London)、加拿大皇后大学(Queen's

University)采用了这种方式,通过在 Pure 的基础上增加开放功能的前端界面,用户可以下载获取科研成果数据。③基于系统间互操作的融合。即机构知识库与科研管理系统独立运行,并通过 API 的方式实现系统间互操作。目前,圣安德鲁斯大学便采用了这种方式,通过使用 Pure 来收集整理相关成果数据,并将能够开放获取的数据推送到机构知识库中,从而源源不断的为机构知识库注入新鲜成果。

北京大学科研管理系统成果子系统,由图书馆负责在 IR 基础上,扩展功能,实现科研管理和机构知识库的统一管理和服务。成果子系统包括 7 个模块(如图 2 所示):登录认证、权限管理、成果提交、成果认领、成果评奖、数据统计、API 接口。原有的 DSpace 系统并不能满足现有需求,因此在 DSpace 的基础上进行了大量二次开发工作。所有这些功能模块都在机构知识库中实现,同时为了更好的满足用户使用习惯,成果认领、成果提交、数据统计等模块也在科研管理系统中实现。北京大学科研管理成果子系统的开发应用,一方面拓展了机构知识库的功能,将其纳入到科研管理过程中,另一方面,通过系统间互操作,最终实现了北京大学机构知识库与科研管理系统的融合,为机构知识库注入了更强的生命力。

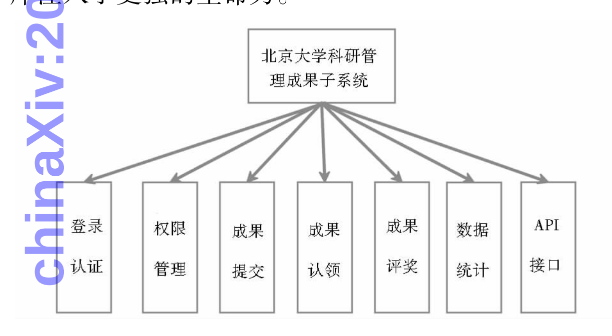


图 2 北京大学科研管理成果子系统模块

3.2.4 从学术成果管理平台到学术资源服务中心 大数据与人工智能快速发展,数据和软件工具日益重要,研究数据作为科学研究的重要成果受到国际学术界和出版领域越来越多的重视。支持管理多种类型的学术资源(包括数据和软件工具),支持研究数据服务,转型成为学术资源服务中心,成为新一代 IR 的重要功能之一。目前,已经有多个机构知识库提供多种方式的数据服务,有建设数据服务平台,收集研究数据并提供服务,北京大学、复旦大学、哈佛大学等多个机构都是采用此种方式, 北京大学还将 IR 成果与成果所用数据通过持久标识符(Handle、DOI)建立关联。也有在 IR 平台上直接扩展收集研究数据并发布服务。D. J. Lee 通过对美国 13 所大型研究型大学的 15 名 IR 管理人员关于研究数据管理服务的访谈,研究 IR 建设中能够提供哪些研究数据服务[50]。布里斯托大学在 IR 平台上,延伸研究数据服务[51]。新一代机构知识库,需要在信息架构上,支持海量数据的管理和服务,提供研究数据服务,为用户提供完整研究生命周期的数据服务。

3.2.5 从海量学术成果数据检索到大数据语义研究支持 大数据和人工智能技术的快速发展,语义搜索技术结合机器学习,提供了更全面和准确的搜索结果。商业数据库已经开始应用与发布相应服务。以 IEEE(ieee.org)与 ip.com 合作发布的专利数据库 InnovationQ Plus(innovationqplus.ieee.org)为例,该数据库使用语义搜索来实现通过概念而不是关键字进行搜索。通过构建语义关系,以等同执行搜索查询时返回的替代单词和短语。通过使用机器学习来提高其概念搜索的准确性。AI 公司 Luminance(luminance.com)和 iManage(imanage.com)使用机器学习和模式识别技术,扫描海量法律文件,分析数据,协助律师分析法律合同[52]。新一代机构知识库,需要研究和应用人工智能和机器学习技术,构建语义搜索,提供数据挖掘和文本挖掘功能,为用户提供研究服务。

3.2.6 从成果存档到新型学术交流社区 学术交流新生态环境下,在线社区交流成为重要的学术交流场景。新一代机构知识库将与用户互动(注释、评论、评述、订阅、主动推送等)服务作为增值服务设重点,力图通过增值服务,使机构知识库成为用户学术交流社区。Facebook、Twitter、微博和微信等社交媒体已经使人们习惯在社交媒体和社区中获取和交流信息。新一代机构知识库在云服务架构和完整学术资源中心基础上,提供学术交流所需的增值服务,将 IR 建设成为人们交流学术信息的新型学术交流社区。

技术、教育与学术交流环境变革,使 IR 的角色与价值发生变化,在学术交流生命周期变革中找到新的角色定位,并发挥作用,成为新一代 IR 建设的目标。

3.2.7 从应用计量指标到建立全新学术评价体系 大量实践和数据已经证明,提供访问统计和引用频次等计量指标,能有效提高机构知识库成果的学术影响力和可见度,众多机构知识库已经提供了多种统计。新一代机构知识库,需要提供更丰富、实时、准确的计量指标数据。通过对数据从总体、机构、研究者、时间等多角度进行统计,生成各层次知识目录。通过对学术资源从多维度或多层面进行逻辑语义关系分析和关联,建立知

识图谱,对机构的知识能力、知识关系、知识资产应用和需求等进行分析和评估。通过对用户使用情况进行分析,建立用户画像。在计量、统计数据、知识图谱、用户画像基础上,建立独立于现有商业出版学术评价数据的全新学术评价体系,重构学术评价生态系统。

4 结语

学术交流快速变革进程中,机构知识库需要重构功能和服务,成为新一代学术交流生态系统的重要基础设施。研究新一代机构知识库的目标、功能、服务和技术,探索新应用,成为当前机构知识库建设的重要内容。中国机构知识库建设,需要抓住这一发展转型机遇,参与到全球学术交流生态系统构建中,成为重要的组成部分,引领世界科研进步。

参考文献:

[1] Opendoar statisticst -an overview of the data held in opendoar[EB/OL]. [2019－03－11]. http://v2. sherpa. ac. uk/view/repository_visualisations/1. html.
[2] Environmental Scan 2017[EB/OL]. [2017－07－01]. http://www. ala. org/acrl/sites/ala. org. acrl/files/content/publications/whitepapers/EnvironmentalScan2017. pdf.
[3] 马建霞. 机构知识库内容建设与服务设计的趋势[J]. 情报理论与实践, 2010(9): 23－27.
[4] 张晓林. 机构知识库的发展趋势与挑战[J]. 数据分析与知识发现, 2014, 30(2): 1－7.
[5] 刘巍, 祝忠明, 张旺强, 等. 基于机构知识库的知识分析及可视化功能实现[J]. 图书与情报, 2016, 36(3): 125－131,137.
[6] 吴志强, 祝忠明, 刘巍, 等. 基于 liresolr 的机构知识库图像检索[J]. 图书馆学研究, 2016, 385(14): 58－63,39.
[7] 吴志强, 祝忠明, 姚晓娜, 等. Cspace 机构知识库影音资源支持能力扩展研究与实践[J]. 数据分析与知识发现, 2017(9): 90－96.
[8] 吴志强, 祝忠明, 刘巍, 等. 机构知识库三维模型检索与展示技术研究与实践[J]. 数据分析与知识发现, 2017(1): 73－80.
[9] Tair 的传承与改变[EB/OL]. [2018－01－01]. http://ntur. lib. ntu. edu. tw/handle/246246/270514#. XAOd8WglI2w.
[10] 崔海媛, 聂华, 罗鹏程, 等. 资助机构开放获取知识库研究与构建－以国家自然科学基金基础研究知识库为例[J]. 图书情报工作, 2017,61(11): 45－54.
[11] Hkust Institutional Repository[EB/OL]. [2018－11－01]. http://repository. ust. hk/ir/.
[12] 香港大学学术库[EB/OL]. [2018－11－01]. http://hub. hku. hk/.
[13] The Polyu Institutional Research Archive (pira)[EB/OL]. [2018－11－01]. http://ira. lib. polyu. edu. hk/.
[14] STERMAN L, BORDA S. Making visualization work for institutional repositories: information visualization as a means to browse electronic theses and dissertations [J]. Journal of librarianship and scholarly communication, 2017, 5(1): 1－17.
[15] COCCIOLO A. Can Web 2.0 enhance community participation in an institutional repository? The case of pocketknowledge at teachers college, columbia university [J]. The journal of academic librarianship, 2010, 36(4): 304－312.
[16] JONES R. Giving birth to next generation repositories [J]. International journal of information management, 2007, 27(3): 154－158.
[17] 从学术典藏库(ir) 到当前科研信息系统(cris) －如何和为何[EB/OL]. [2017－11－01]. https://core. ac. uk/download/pdf/38034688. pdf.
[18] Reshaping library support for research at King's College London [EB/OL]. [2017－11－01]. http://www. unica-network. eu/sites/default/files/GB% 20and% 20NW% 20UNICA% 20Brussels% 202012. pdf.
[19] DE CASTRO P, SHEARER K, SUMMANN F. The gradual merging of repository and cris solutions to meet institutional research information management requirements [J]. Procedia computer science, 2014, 33(2014): 39－46.
[20] FINA F, PROVEN J. Using a cris to support communication of research: mapping the publication cycle to deposit workflows for data and publications [J]. Procedia computer science, 2017, 106(2017): 232－238.
[21] Technical recommendations for next generation repositories[EB/OL]. [2018－12－20]. https://www. coar-repositories. org/news-media/technical-recommendations-for-next-generation-repositories/.
[22] Next generation repositories[EB/OL]. [2018－11－15]. https://www. coar-repositories. org/activities/advocacy-leadership/working-group-next-generation-repositories/.
[23] Science Europe[EB/OL]. [2018－11－01]. https://www. scienceeurope. org/coalition-s/.
[24] Wellcome is updating its open access policy[EB/OL]. [2018－11－08]. https://wellcome. ac. uk/news/wellcome-updating-its-open-access-policy.
[25] 张晓林. 让所有科研论文免费阅读,中国机构明确力挺开放获取[EB/OL]. [2018－12－20]. http://zhishifenzi. com/depth/depth/4778. html.
[26] Guidance on the implementation of Plan S[EB/OL]. [2018－12－20]. https://www. coalition-s. org/feedback/.
[27] Coar's response to draft implementation requirements in Plan S [EB/OL]. [2018－12－20]. https://www. coar-repositories. org/news-media/coars-response-to-draft-implementation-requirements-in-plan-s/.
[28] Impact of Plan S Implementation guidelines on dspace repositories [EB/OL]. [2018－12－20]. https://www. atmire. com/articles/detail/impact-of-plan-s-implementation-guidelines-on-dspace-repositories.
[29] CROW R. The case for institutional repositories: a sparc position paper, 2002[EB/OL]. [2018－12－21]. https://uta-ir. tdl. org/uta-ir/bitstream/handle/10106/24350/Case% 20for% 20IRs_SPARC. pdf? sequence=1.

[30] Signposting the scholarly web[EB/OL]. [2018－11－29]. http://signposting. org/.

[31] Activity streams 2.0[EB/OL]. [2018－11－28]. https://www. w3. org/TR/activitystreams-core/.

[32] Open annotation community group[EB/OL]. [2018－11－28]. https://www. w3. org/community/openannotation/.

[33] International image interoperability framework[EB/OL]. [2018－11－08]. https://iiif. io/.

[34] Ipfs is the distributed web[EB/OL]. [2018－11－01]. https://ipfs. io/.

[35] Resourcesync framework specification[EB/OL]. [2018－11－01]. http://www. openarchives. org/rs/toc.

[36] About SWORD[EB/OL]. [2018－11－01]. http://swordapp. org/about/.

[37] What are sitemaps? [EB/OL]. [2018－11－01]. https://www. sitemaps. org/.

[38] Linked data notifications[EB/OL]. [2018－11－01]. https://www. w3. org/TR/ldn/.

[39] Resourcesync framework specification - change notification[EB/OL]. [2018－11－01]. http://www. openarchives. org/rs/notification/1. 0. 1/notification.

[40] Webmention[EB/OL]. [2018－11－01]. https://www. w3. org/TR/webmention/.

[41] Websub[EB/OL]. [2018－11－01]. https://www. w3. org/TR/websub/.

[42] Orcid[EB/OL]. [2018－11－01]. https://orcid. org/.

[43] Webid 1.0[EB/OL]. [2018－11－01]. https://www. w3. org/2005/Incubator/webid/spec/identity/.

[44] Signing http messages[EB/OL]. [2018－11－01]. https://datatracker. ietf. org/doc/draft-cavage-http-signatures/.

[45] Openid connect[EB/OL]. [2018－11－01]. https://openid. net/connect/.

[46] Webid authentication over tls[EB/OL]. [2018－11－01]. https://www. w3. org/2005/Incubator/webid/spec/tls/.

[47] Counter [EB/OL]. [2018－11－01]. https://www. project-counter. org/.

[48] Http etag[EB/OL]. [2018－11－01]. https://en. wikipedia. org/wiki/HTTP_ETag.

[49] Public access plan: today's data, tomorrow's discoveries: Increasing access to the results of research funded by the national science foundation[EB/OL]. [2018－10－01]. https://www. nsf. gov/news/special_reports/public_access/#.

[50] LEE D J, STVILIA B. Practices of research data curation in institutional repositories: a qualitative view from repository staff [J]. PloS one, 2017, 12(3): e0173987.

[51] 唐凤. 面向科研数据管理的科研信息系统与机构知识库的链接研究 [J]. 情报理论与实践, 2018, 41(2): 73－76.

[52] OJALA M. Big data and ai: Technology, transparency, and trust [EB/OL]. [2019－01－01]. http://www. infotoday. com/cilmag/dec18/Ojala--Big-Data-and-AI-Technology-Transparency-and-Trust. shtml.

作者贡献说明：

崔海媛：设计论文框架、内容撰写与修改论文；
孙超：关键技术调研与修改论文；
罗鹏程：关键技术调研与修改论文。

Key Technology and Trends of the Next Generation Repositories

Cui Haiyuan Sun Chao Luo Pengcheng
Peking University Library, Beijing 100871

Abstract: [**Purpose/significance**] By investigating the research status and service demand of the new generation knowledge at home and abroad, this paper analyzes the key technologies and functional characteristics of the next generation repositories. It puts forward the key trends of institutional repositories, and provides some suggestions for the development of next generation repositories. [**Method/process**] Through literature research, this paper summarized the development trend of institutional repository technology and function. Then, it introduced 11 key technologies, standards and protocols of the new generation repositories. By researching and designing the functions, it analyzed the development of next generation repositories in many aspects such as framework, functions and service target. [**Result/conclusion**] The paper generalizes the development trends of next generation repositories as followings: ①from institutional repository to institutional information infrastructure. ②from self-archiving to automated submission or ingestion. ③ from independent platforms to current research information system(CRIS). ④ from scholarly outputs management platform to academic resources service center. ⑤from information retrieval to supporting big data and semantic retrieval. ⑥from academic archives to academic scholarly communities. ⑦from providing metrics and altmetrics services to creating a new system of academic evaluation.

Keywords: institutional repository new generation repository next generation repository key technology development trend